The Masked Advantage: Uncovering Local-Language Access to Cultural Knowledge in LLMs

Yang Zhang[1†], Xiao Fei[1], Amr Mohamed[1,2], Sarah Almeida Carneiro[1], Mersin Konomi[2], Mingmeng Geng[3], Ahmed Asaad[4], Guokan Shang[2], Michalis Vazirgiannis[1,2†]

[1]École Polytechnique, [2]MBZUAI, [3]ENS-PSL, [4]Durham University

[†]Correspondence: yang.zhang@polytechnique.edu, mvazirg@lix.polytechnique.fr

Abstract

Large language models are increasingly used to answer culturally grounded questions across languages, yet it remains unclear whether local cultural knowledge is better accessed through English or the local language. Existing evaluations face two key limitations: many rely on parallel template-based questions that may not reflect how cultural knowledge naturally appears, and raw accuracy conflates general language proficiency with language-conditioned knowledge access. We address these issues with a controlled framework built on real-world cultural questions collected from regional benchmarks and local sources. By crossing question type (culture-agnostic vs. culture-specific) with query language (English vs. local language), and estimating ability with a shared 1PL item response theory model, we separate proficiency from localized knowledge access. Across 13 locales and roughly 80 models, we find a consistent English advantage on culture-agnostic questions, indicating stronger English proficiency. However, after accounting for this proficiency gap, local languages show a positive knowledge-access advantage in nearly all locale–model settings. This advantage is often masked in raw accuracy but becomes more visible for frontier, regionally aligned, or language-adapted models. Our results suggest that weaker local-language performance does not necessarily imply weaker cultural knowledge; rather, local cultural knowledge may be more accessible through the local language but hidden by limited language proficiency.

1 Introduction

Large language models are increasingly regarded as multilingual systems, capable of answering questions written in different languages and concerning history, law, customs, and everyday life across diverse cultural contexts. However, the relevant cultural knowledge is often unevenly distributed across languages: some content is produced and discussed primarily in a local language, while training data and evaluation benchmarks remain dominated by English — the global language. This raises two linked questions: *Q1: which query language provides better access to local cultural knowledge?* and *Q2: whether that knowledge is preferentially accessed through the language in which it is grounded?*

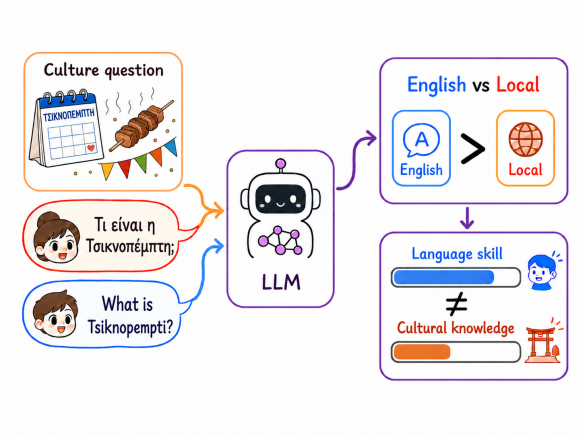

Figure 1: For cultural questions, local knowledge may be masked by weaker local-language proficiency.

Prior work has yielded mixed findings. Several studies suggest that LLMs remain strongly English-centric: knowledge transfers asymmetrically across languages, and translating into English can be a competitive querying strategy, especially for lower-resource languages (Yin et al., 2022; Liu et al., 2025; Romanou et al., 2025). Another line reports that local-language prompting can help when the query language aligns with the cultural context, and that different languages may activate different knowledge in the same model (Ying et al., 2025; Agarwal et al., 2025; Jain and Chopra, 2026; Myung et al., 2024).

We argue this disagreement stems partly from how cultural knowledge is measured. Many cultural benchmarks build parallel questions from

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 40007–40023
October 24-29, 2026 ©2026 Association for Computational Linguistics

human-defined templates across languages, which may not reflect how local knowledge naturally appears in each language (Ying et al., 2025).[1] Even with faithful benchmarks, most cross-lingual comparisons rely on raw accuracy, which conflates language proficiency with access to language-conditioned knowledge—two effects that may work in opposite directions. A model may score higher in English simply because it understands English better, while a local language may better activate culturally grounded knowledge yet still yield lower accuracy due to weaker proficiency. A raw English-versus-local gap therefore reports only the net sign of these factors, leaving open which language truly provides better access to local knowledge.

To separate these effects, we compare English and local-language performance on both culture-agnostic and culture-specific questions. The culture-agnostic gap estimates general language proficiency (GLOBALGAP), while the culture-specific gap captures both proficiency and cultural knowledge access (LOCALGAP). Their difference gives KNOWLEDGEGAP, which isolates language-conditioned knowledge access.

However, because culture-agnostic and culture-specific items are drawn from different question pools of different intrinsic difficulties, raw accuracy gaps across the two sets are not directly comparable. A lower accuracy on culture-specific questions may simply reflect harder items rather than weaker knowledge access. Therefore, we estimate all gaps with a shared 1PL IRT model (Lalor et al., 2016; Martínez-Plumed et al., 2019), which jointly estimates item difficulty and model ability on a common scale, enabling the comparison by subtraction. To our knowledge, no prior work combines item-response modeling with a difference-in-differences design in this way.

We apply this framework across 13 locales and roughly 80 models, using culture-specific items from established regional benchmarks rather than synthetic templates. We deliberately build no new dataset: culturally grounded items must be natively authored in the local language, so we source them from published regional benchmarks worldwide rather than translating English ones. Our results support the masked advantage hypothesis: models show a consistent English proficiency advantage on culture-agnostic questions, while local-language performance on culture-specific questions is mixed. After controlling for proficiency, however, KNOWLEDGEGAP is positive in nearly every locale–model cell, including many where LOCALGAP is negative. This suggests that local languages often provide better access to cultural knowledge, but this advantage is hidden by weaker local-language proficiency.

Our contributions are threefold:

1. We introduce a measurement framework that separates language proficiency from language-conditioned cultural knowledge access, casting all three gaps as ability contrasts under a single 1PL (Rasch) IRT model that accounts for difficulty differences between culture-agnostic and culture-specific items.

2. We conduct a large-scale empirical study across 13 local languages and multiple model categories using natively sourced items from established benchmarks rather than synthetic probes.

3. We provide evidence for a masked local-language knowledge advantage that is present in nearly all settings but surfaces in raw accuracy only when proficiency does not work against it. Weaker local-language performance therefore appears to stem largely from limited proficiency in that language rather than from missing cultural knowledge itself, making this largely a multilingual-access problem rather than one of missing cultural knowledge, with implications for how multilingual LLMs are evaluated and deployed.

2 Related Work

Multilingual, Cultural, and Regional Evaluation Benchmarks. A growing body of benchmarks evaluates LLMs beyond English. Translation-based suites such as Global-MMLU (Singh et al., 2025) show that multilingual evaluation can still inherit English-centric cultural assumptions. While native or regional benchmarks such as INCLUDE (Romanou et al., 2025), CMMLU (Li et al., 2024), ArabicMMLU (Koto et al., 2024), KMMLU (Son et al., 2025), MILU (Verma et al., 2025), BertaQA (Etxaniz et al., 2024), and GreekMMLU (Zhang et al., 2026) collect questions from local exams, institutions, and knowledge sources. In parallel, cultural benchmarks such as BLEnD (Myung et al., 2024) extend evaluation from academic knowledge

[1] For instance, a template may ask what people in China eat at Christmas, a holiday not traditionally celebrated there.

to everyday culture but with a predefined template. There are also works evaluating LLMs under code-switched tests, revealing mixed effects of English and local language (Mohamed et al., 2025). These benchmarks establish that a language gap exists, but report raw accuracy without separating language proficiency from knowledge access; our framework supplies that decomposition.

Disentangling Language from Knowledge. Closest to our setting, Ying et al. (2025) separates language from cultural context and finds that local-language prompts generally improve cultural knowledge access, but it relies on parallel template-based cultural questions, which may not capture how localized knowledge naturally appears in native data. Building on this direction, Agarwal et al. (2025) further asks whether some knowledge is more accessible in a specific language than in English, while Jain and Chopra (2026) argues that changing the query language can also shift the cultural assumptions activated by the model. Etxaniz et al. (2024) show the same effect for Basque. We instead collected real-world cultural questions and cast the problem as a difference-in-differences of model ability estimated by a Bayesian item-response model (Lalor et al., 2016; Rodriguez et al., 2021). This lets us recover a knowledge advantage that raw accuracy masks.

IRT for LLM MCQ Benchmarking The community has witnessed an increasing number of works applying Item Response Theory (IRT) to LLM Multiple-Choice Question (MCQ) benchmarking, enabling analysis of LLM ability separately from item difficulty. Both tinyBenchmarks (Polo et al., 2024) and metabench (Kipnis et al., 2025) use IRT on LLM responses across MCQs to distill benchmarks into much smaller sets, while variants such as PSN-IRT (Zhou et al., 2026) and β^3-IRT (Chen et al., 2019) have previously been proposed to extend the capacity of vanilla IRT for more detailed analysis of item difficulty and discrimination. Some recent studies further examine the fitting quality of IRT models with Computerized Adaptive Testing (CAT) (Zhuang et al., 2023a; Li et al., 2025). These works rank or compress benchmarks rather than contrast abilities across query languages. However, very few of them have carefully examined the translation and scale invariance problem of logistic IRT models (Zhuang et al., 2023b; Li et al., 2025), which is crucial in our particular case of estimating independent model ability on different subsets of the benchmark.

3 Methodology

3.1 Problem Formulation

Let j be a language model. We study two subsets of problems $s \in \{\text{CA}, \text{CS}\}$, where culture-specific questions Q_{CS} target a region with local language k_{local}, alongside culture-agnostic questions Q_{CA} that require no local knowledge. Each subset is posed both in k_{local} and in English $k = \text{en}$, so that the effect of query language is measured within a fixed question pool.

We define all gaps in terms of model ability θ_{jsk} rather than raw accuracy, because comparing accuracy across question sets confounds the effect of interest with differences in their difficulty—a concern for our central contrast, which compares the disjoint pools Q_{CS} and Q_{CA}. We estimate θ_{jsk}, the ability of LLM j on question set s posed in language k, with an Item-Response Model that places item difficulty and model ability on a common scale (§3.2).

Our first goal is to examine the LOCALGAP, the ability difference between querying in k_{local} and in English on culture-specific questions,

$$\text{LOCALGAP}_j = \theta_{j,\text{CS},k_{local}} - \theta_{j,\text{CS},\text{en}}. \quad (1)$$

Its sign alone, however, does not reveal *why* one language wins, because it conflates *language proficiency*—general ability in k_{local} versus English—with *language-conditioned knowledge access*—whether culturally specific knowledge is preferentially encoded in or retrievable through k_{local}. To isolate proficiency, we measure the same contrast on culture-agnostic content, the GLOBALGAP,

$$\text{GLOBALGAP}_j = \theta_{j,\text{CA},k_{local}} - \theta_{j,\text{CA},\text{en}}, \quad (2)$$

on which no local-knowledge advantage is expected.

Our second goal is to separate the two effects. Subtracting the GlobalGap from the LocalGap cancels proficiency and isolates knowledge access; we call this difference-in-differences contrast the KNOWLEDGEGAP,

$$\text{KNOWLEDGEGAP} = \text{LOCALGAP} - \text{GLOBALGAP} \quad (3)$$

A positive KNOWLEDGEGAP indicates a local-language advantage on culture-specific content beyond what proficiency explains.

3.2 1PL-IRT Model

In order to estimate model ability and task difficulty separately, we use a One-Parameter Logistic Item Response Theory (1PL-IRT) model, equivalent to the Rasch model (Bond et al., 2015). making abilities comparable across benchmarks of varying difficulty. It models the probability that examinee j answers item i correctly through the examinee's ability θ_j and the item difficulty b_i. In our setting, each logical item is prompted in multiple languages, and we want difficulty to reflect the problem rather than the prompt language.

Assumption 1 (Translation Equivalence)**.** *An underlying MCQ item has the same intrinsic difficulty across the languages it is translated into.*

Definition 2 (Multi-Facet Problem)**.** A single logical MCQ item $i \in I$ is prompted through a facet $k \in K$, i.e. a translated language. By Assumption 1, the difficulty b_i is one scalar shared across all facets k.

The residual variation in correctness across languages is then absorbed by an ability specific to each subset $s \in \{\text{CS}, \text{CA}\}$ and language k. Writing $Y_{ijsk} \in \{0, 1\}$ for the response of examinee j to item i in subset s and language k:

Definition 3 (Facet-Conditioned 1PL-IRT Model)**.**

$$P(Y_{ijsk} = 1 \mid \theta_{jsk}, b_i) := \sigma(\theta_{jsk} - b_i), \quad (4)$$

with σ the logistic function and θ_{jsk} the ability of j on subset s in language k.

Assuming local independence, the model factorizes over items. Since the facet enters through θ rather than b, the within-item log-odds difference across two languages cancels difficulty,

$$L_{ijsk} - L_{ijsk'} = \theta_{jsk} - \theta_{jsk'}, \quad (5)$$

making language contrasts difficulty-free — the property our gap metrics rely on. The full specification appears in Appendix D.

3.3 Dataset Construction and Languages

We cover 13 languages (locales) spanning diverse resource levels, scripts, and regions: Chinese, Arabic, Greek, Hindi, Korean, French, Japanese, Bengali, Dutch, Nepali, Persian, Polish, and Telugu. For each locale, we build three matched subsets: culture-agnostic items (CA), culture-specific items in the local language (CS), and their English counterparts (CS_{EN}). Each language is associated with a resource level described in (Joshi et al., 2020). Appendix A reports per-locale statistics.

Culture-agnostic (CA). We draw CA items from Global-MMLU, which provides high-quality parallel translated questions across languages and human-assigned culture-agnostic labels; for each locale, we select the CA items in the corresponding language.

Culture-specific (CS). CS items are natively authored in the local language, not translated from English. They come from national exams and local knowledge sources compiled by regional benchmarks: CMMLU (Li et al., 2024), ArabicMMLU (Koto et al., 2024), GreekMMLU (Zhang et al., 2026), MILU (Verma et al., 2025), and KMMLU (Son et al., 2025), with the remaining locales drawn from INCLUDE. From these we manually select subsets that are potentially cultural-related, and then ask the native speakers to select only genuinely culture-specific items, as described in Appendix B. We deliberately build no new dataset: genuinely cultural items must be written natively, which templated or translated construction cannot provide at this scale.

English counterparts (CS_{EN}). We translate each CS item into English with Claude Sonnet 4. We hired native speakers to assess every item, rating translation quality and labeling whether the question is culturally specific; we retain only high-quality translations with genuine cultural content. CS and CS_{EN} thus contain the same logical questions in two query languages, isolating the effect of query language on cultural knowledge access. Annotation details are in Appendix B.

4 Experiments

4.1 Models

We evaluate roughly 80 models in three groups, testing both base and instruction-tuned variants where available. Appendix A lists each models with its parameter count and country of origin.

Closed-source frontier. The proprietary models GPT-5.4, GPT-5.4-mini (OpenAI, 2026), Gemini-3-Flash (Google DeepMind, 2025), and Claude-Sonnet-4.6 (Anthropic, 2026).

Multilingual pretrained. Open-weight models pretrained from scratch on large, multilingual corpora, spanning a range of countries of origin: Llama (Grattafiori et al., 2024) and Gemma (Team et al., 2024, 2025b; Google DeepMind, 2026) (USA), Qwen (Qwen et al., 2025; Yang et al.,

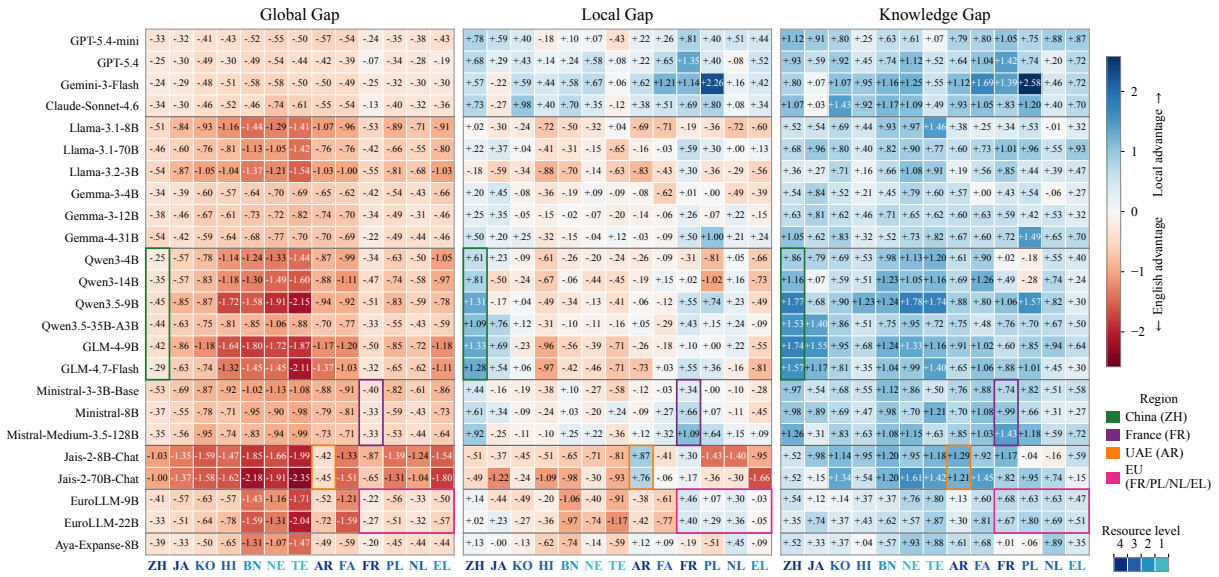

Figure 2: GlobalGap, LocalGap, and KnowledgeGap across models and locales, showing a consistent English proficiency advantage but a near-universal local-language knowledge advantage.

2025; Team, 2026) and GLM (Glm et al., 2024; Team et al., 2025a) (China), Mistral (Rastogi et al., 2025; Liu et al., 2026) (France), Jais-2 (Sengupta et al., 2023) (UAE), EuroLLM (Martins et al., 2024) (EU), and Aya-Expanse (Dang et al., 2024) (Canada).

Regionally-adapted. Models obtained by continued pretraining or fine-tuning of a global base on local-language and cultural data, rather than pretrained from scratch: AceGPT-v2 (Arabic) (Liang et al., 2024), Llama-Krikri (Roussis et al., 2025) and Meltemi (Voukoutis et al., 2024) (Greek), PersianMind (Persian) (Rostami et al., 2024), and Bielik (Polish) (Ociepa et al., 2024).

Model filtering. Our decomposition assumes each model has enough multilingual competence for the gaps to be meaningful. We therefore retain only models whose accuracy exceeds 0.35 on every subset, keeping the estimation within a regime where the model demonstrably engages with each condition.

4.2 Prompts and Inference

Prompt templates. Each cell uses a fixed template rendered entirely in the query language, so no cross-lingual cues leak in. CA questions use a `Question: {question+options} Answer:` format; CS questions prepend a locale cue, `In {country}, Question: {question+options} Answer:`, to ground them in local context. Each template is then translated into the query language; CS_EN uses the same CS template in English without translation. We apply no chat template and no reasoning/chain-of-thought prompting, consistent with our log-likelihood scoring. For closed-source models, we use free-generation and extract the answer option.

Inference. All open-weight models run in bfloat16 without quantization at temperature 0 with a fixed seed. Smaller models run on NVIDIA A6000 GPUs and larger models on A100 80GB.

4.3 IRT Fitting

At our benchmarking scale, classical estimators such as Newton–Raphson or Expectation–Maximization are less practical. We instead frame the facet-conditioned 1PL model as a continuous optimization problem: treating the ability matrix $\boldsymbol{\Theta}$ and difficulty vector $\boldsymbol{B}$ as trainable embeddings, we recover the maximum-likelihood estimates by minimizing the negative log-likelihood, which coincides with the binary cross-entropy loss, using the Adam optimizer (Kingma and Ba, 2017) with mini-batch SGD, more details are in Appendix C.

The 1PL likelihood is non-identifiable: since σ depends only on $\theta_{jsk} - b_i$, adding a constant to all abilities and all difficulties leaves every probability unchanged. To guarantee a unique solution we anchor the latent scale by constraining the item difficulties to mean zero, applying the projection $b_i \leftarrow b_i - \bar{b}$ and $\theta_{jsk} \leftarrow \theta_{jsk} - \bar{b}$ (with $\bar{b} = \frac{1}{|I|}\sum_{i \in I} b_i$) after each update. This fixes

the origin of the scale and gives a unique, interpretable solution. The full derivation appears in Appendix D.2.

5 Results

Figure 2 reports a representative subset of model results. For readability, we select models that are generally pretrained from scratch and cover diverse model families, regions of origin, and parameter scales. Regionally adapted models are discussed separately in Section 6. Full results for all tested models are provided in Appendix E.

5.1 A Universal English Proficiency Advantage

Across the full grid, GLOBALGAP is negative in almost every cell (Figure 2, left panel): for all models, on all 13 languages, the estimated ability on culture-agnostic questions is higher under English queries than under local-language queries. The cell-level mean is -0.79, and almost no cell is positive. This pattern holds even for the highest-resource non-English languages in our set, including Chinese (-0.44) and French (-0.37), which still favor English on the CA subset. We read this as primarily a proficiency baseline: largely independent of cultural content, English yields higher estimated ability for these models, and the effect is consistent enough that we treat it as a background against which the other two gaps should be interpreted.

5.2 The Local Language Tends to Win With High Resources or Matching Origin

Unlike GLOBALGAP, LOCALGAP is mixed (Figure 2, center panel). The positive, local-favoring cells concentrate on the highest-resource languages: Chinese has the most positive column mean ($+0.48$), followed by French ($+0.42$), the two clearest cases where querying in the local language yields higher estimated ability on culture-specific questions than translating to English. For most other languages LOCALGAP is negative or near zero, so on the CS ability contrast translating to English is generally the better choice.

A second factor is model origin. Among models built in a given region, the home-language LOCALGAP tends to be markedly higher than for other models on the same language: $+0.79$ (Chinese-origin models on Chinese), $+0.94$ (UAE-origin on Arabic), and $+0.32$ (French-origin on French) (Table 2). The Arabic case is notable because Arabic's overall column mean is slightly negative, so here the advantage appears to come from origin rather than resource level. The local language thus tends to win when it is resource-rich, matched to the model's region of origin, or both.

5.3 The Knowledge Advantage Is Universal but Driven by Origin, Not Resources

KNOWLEDGEGAP, the knowledge-access component (Figure 2, right panel), is positive in 98% of cells: the knowledge component favors the local language almost everywhere, including cells where the corresponding LOCALGAP is negative.

This dissociation is the point of the decomposition: once proficiency is partialled out, the local language carries a positive knowledge advantage.

Unlike the two performance gaps, KNOWLEDGEGAP is not explained by resource level. What does track it is model origin: in-region models carry a larger home-language KNOWLEDGEGAP than other models ($+0.70$ ZH, $+0.60$ AR, $+0.30$ FR; Table 2), suggesting the advantage is broadly present across models but is deepened by targeted regional training rather than corpus scale.

5.4 Performance Gaps Scale with Resource Level

The English advantage is not uniform across languages but varies with resource level. Table 1 groups 13 languages into four ordinal resource levels and reports the mean of each gap. Both GLOBALGAP and LOCALGAP become less negative as resource level rises, moving from -1.22 to -0.54 and from -0.27 to $+0.22$ respectively across the 13 per-language means, resource level correlates with both (GLOBALGAP: Pearson $r = +0.84$, $p < 0.001$; LOCALGAP: $r = +0.69$, $p = 0.009$). The KNOWLEDGEGAP shows no comparable trend ($r = -0.30$), a dissociation as discussed in Section 5.3. The proficiency penalty thus shrinks as a language becomes higher-resource, consistent with GLOBALGAP being driven primarily by how well the model handles the language itself.

5.5 Closed-Source Models Surface the Local-Language Advantage

The four proprietary frontier models—GPT-5.4-mini, GPT-5.4, Gemini-3-Flash, and Claude-Sonnet-4.6 answer most culture-specific questions better in the local language than in English, with positive LOCALGAP in most cells alongside positive KNOWLEDGEGAP in nearly all of them. Be-

Level	Languages	Global	Local	Know.
1	TE, NE	−1.22	−0.27	0.95
2	BN, EL	−0.95	−0.26	0.70
3	HI, FA, KO, PL, NL	−0.76	−0.07	0.68
4	AR, JA, ZH, FR	−0.54	+0.22	0.77

Table 1: Mean gaps by resource level, averaged over models within each level group. Both GLOBALGAP and LOCALGAP rise (become less negative) with resource level; KNOWLEDGEGAP does not show a clear monotonic trend.

cause these models carry the smallest proficiency penalty (GLOBALGAP among the least negative in the grid), the local-language advantage is no longer masked and surfaces directly in the uncorrected contrast—the same mechanism behind the Chinese result, here driven by model capability rather than resource level. This also indicates the culture-specific items genuinely require local knowledge.

6 Ablations

6.1 Both Gaps Grow with Model Scale

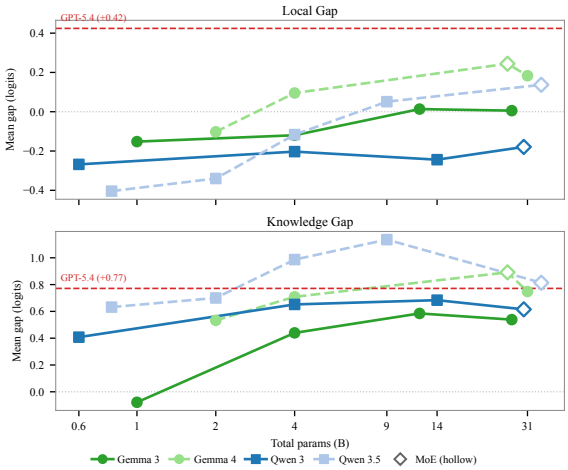

Figure 3: LOCALGAP (top) and KNOWLEDGEGAP (bottom) versus model size for four model families. The red dashed line marks the frontier reference GPT-5.4.

To probe how the decomposition behaves with scale, we plot LOCALGAP and KNOWLEDGEGAP against total parameter count for four model families (Figure 3). Both gaps tend to grow with size. The trend is clearest for LOCALGAP, which rises within every family—most steeply for the newer Gemma 4 and Qwen 3.5, gently for Gemma 3, and roughly flat for Qwen 3—consistent with scale gradually reducing the proficiency penalty that masks the local-language advantage.

KNOWLEDGEGAP also increases overall but not monotonically: some families peak at intermediate sizes and then flatten or decline, most visibly Qwen 3.5 near 9B. The MoE checkpoints move inconsistently relative to dense models of similar size, so we draw no general conclusion about MoE. The two metrics thus scale differently—LOCALGAP rising steadily, KNOWLEDGEGAP rising then leveling—but both persist across the full size range, mostly below the GPT-5.4 reference, rather than being tied to any particular scale.

6.2 Language Adaptation Helps LocalGap but Not Necessarily KnowledgeGap

We ask whether the masked advantage persists when a model is explicitly adapted to a single target language. If the LOCALGAP penalty is mainly a proficiency deficit, language-specific supervised fine-tuning (SFT) should reduce or reverse it, whereas the KNOWLEDGEGAP—which reflects where cultural knowledge is accessible rather than surface proficiency—need not move. We evaluate five publicly available models adapted toward a non-English language, each reported on its adapted language and compared against a *per-language* baseline: the average of the general pretrained models on that same language.

Table 3 shows two consistent effects and one that divides by language. LOCALGAP turns positive for all five models, above a baseline that is itself negative or near zero, and GLOBALGAP rises toward zero for every model but stays negative; adaptation thus narrows the proficiency gap to English without closing it, consistent with proficiency being what SFT most readily moves. KNOWLEDGEGAP, by contrast, does not move uniformly: the European models gain over their baseline while the Persian and Arabic models fall slightly below it. This split suggests that SFT adds target-language cultural knowledge for Greek and Polish, but for Persian and Arabic mainly improves linguistic alignment, with little knowledge advantage beyond the same-language general models.

6.3 Latent Ability θ Recovers Raw Accuracy

Because θ is estimated from observed responses, we verify it faithfully reflects raw accuracy, so that ability contrasts can be read as difficulty-adjusted accuracy differences. Across all subsets, the mean Pearson correlation between accuracy and θ is 0.996, ranging narrowly from 0.982 (Greek CS) to 0.999 (Arabic CS), and the Spearman correla-

Region → Lang	LocalGap In	LocalGap Out	LocalGap Δ	KnowledgeGap In	KnowledgeGap Out	KnowledgeGap Δ
China → ZH (n=6)	+1.07	+0.28	+0.79	+1.44	+0.74	+0.70
UAE → AR (n=2)	+0.82	−0.12	+0.94	+1.25	+0.65	+0.60
France → FR (n=3)	+0.70	+0.38	+0.32	+1.05	+0.75	+0.30
Europe → FR/PL/NL/EL (n=2)	+0.23	+0.05	+0.18	+0.64	+0.64	−0.01

Table 2: Home-language gaps for in-region models versus all other models, for LocalGap and KnowledgeGap. In-region models show a larger LocalGap on their home language in every case. The KnowledgeGap boost holds for the single-language regions (China, UAE, France).

Model	Lg	Global Adapt.	Global Δ	Local Adapt.	Local Δ	Know. Adapt.	Know. Δ
Krikri-8B	EL	−0.44	+0.33	+0.39	+0.65	+0.83	+0.32
Meltemi-7B	EL	−0.47	+0.30	+0.48	+0.74	+0.94	+0.44
Bielik-Minitron-7B	PL	−0.41	+0.25	+0.75	+0.64	+1.16	+0.40
PersianMind	FA	−0.55	+0.34	+0.14	+0.18	+0.69	−0.16
AceGPT-v2-8B	AR	−0.23	+0.52	+0.31	+0.36	+0.54	−0.16
Mean (Adapt.)		−0.42	+0.35	+0.41	+0.52	+0.83	+0.17

Table 3: Language-adapted models evaluated on their target languages *Adapt.* denotes the adapted model's gap value. Δ denotes the difference between the adapted model and the per-language baseline, defined as the mean gap of general-purpose models evaluated on the same language.

tion is 0.9999, indicating that θ recovers the examinee rank order almost exactly. This is expected: within a subset, θ is by construction a monotonic re-expression of accuracy on a continuous log-odds scale. As a result, the within-component GlobalGap and LocalGap closely track the corresponding local-minus-English accuracy differences; the IRT scale adds value by putting them on the same difficulty-adjusted footing as the cross-component KnowledgeGap, a comparison raw accuracy cannot support.

6.4 Selection of the IRT Model

We ablate two orthogonal choices in a 2×2 grid: the item-response representation (binary vs. a Continuous Response Model, CRM, which scores each item by the normalized probability mass over option tokens rather than a hard label) and the inference algorithm (maximum likelihood, MLE, vs. the No-U-Turn Sampler, NUTS); both are defined in Appendix D.3. All four combinations are evaluated under 5-fold out-of-distribution cross-validation on held-out responses, scored by predictive log-likelihood (BCE), calibration (Brier), and discrimination (AUC, accuracy) (Table 4).

The binary representation outperforms CRM under both inference methods, and MLE matches or slightly exceeds NUTS at far lower cost—likely because exact posterior sampling is difficult in this high-dimensional parameter space. The CRM gap is consistent with soft option-token probabilities being noisy and poorly calibrated, so thresholding to binary correctness yields a more reliable ability estimate. We therefore adopt `binary+mle`.

Variant	BCE↓	SD	Brier↓	AUC↑	Acc↑
binary+mle	**0.513**	9e-4	**0.169**	**0.830**	**0.752**
binary+nuts	0.515	1e-3	0.170	0.829	0.748
crm+mle	0.537	5e-4	0.178	0.816	0.746
crm+nuts	0.553	8e-4	0.182	0.821	0.746

Table 4: Cross-validation performance across IRT variants. BCE and Brier measure predictive likelihood and calibration; AUC and accuracy measure discrimination on held-out responses. A standard deviation (SD) is reported along with BCE for different folds.

6.5 The Origin Effect Is Subject-Concentrated

Section 5.3 showed that the knowledge advantage tracks model origin. To ask where in the item space it sits, we group culture-specific items into five subject groups and compute the performance improvement of home-region models over all other models on the same language.

Table 5 shows the effect is present in every group but concentrated rather than uniform. History gains most for Chinese (+53%) and Religion & Culture for Arabic (+38%), both areas whose knowledge

Subject group	Chinese	Arabic
History	$+53\%$	$+28\%$
Religion & Culture	$+35\%$	$+38\%$
Law	$+44\%$	$+21\%$
Professional	$+52\%$	$+21\%$
Language	$+41\%$	$+20\%$

Table 5: Relative performance improvement of home-region models over all other models, by subject group, on culture-specific items in the home language.

is written and discussed predominantly in the local language. That every group improves, rather than one, indicates the effect is not an artefact of a single topic or benchmark source.

6.6 English Translation Does Not Rescue Hard Items

If the local-language penalty were purely a language barrier, translating to English should help most where local-language performance is worst. It does not. On the hard tail — items solved by at most 15% of models under local-language prompting, and 8.8% of all culture-specific items, so the benchmark is not dominated by unanswerable questions — translation recovers only 0.2% of items.

English thus helps in the middle of the difficulty distribution, not at the tail, as the masked-advantage account predicts: a genuinely hard item reflects knowledge the model does not hold, which no query language can retrieve, whereas mid-difficulty items hold knowledge that the proficiency penalty conceals.

6.7 Answer-Position Bias and Its Effect on the Gaps

When a model cannot parse a prompt, it may drift toward the first option rather than answer from knowledge. Measuring over-selection of the first option relative to chance, we find it largest on culture-specific local prompts ($+3.62$ pp) and shrinking monotonically as prompts become easier to read ($+2.54$ CS-English, $+2.04$ CA-local, $+1.14$ CA-English), in a content-independent $A > B > C > D$ gradient. Part of the local-language penalty is therefore parsing failure rather than absent knowledge, supporting the reading of GLOBALGAP as a proficiency term. Because the bias loads on culture-agnostic and culture-specific items alike, it largely cancels in the difference-in-differences contrast and leaves KNOWLEDGEGAP essentially unaffected.

7 Conclusion

We introduced a controlled framework that combines a difference-in-differences design with item-response modeling to study language-conditioned access to localized knowledge in LLMs. By crossing question type, culture-agnostic versus culture-specific, with query language, English versus local, we separate general language proficiency from localized knowledge access. Specifically, we estimate GLOBALGAP as the proficiency gap, LOCALGAP as the combined effect on culture-specific questions, and KNOWLEDGEGAP as their difference on a shared, difficulty-adjusted 1PL IRT scale.

Across 13 locales and roughly 80 models, LOCALGAP is shaped by language resource level, model origin, and language adaptation, becoming positive mainly when these factors favor the local language. In contrast, KNOWLEDGEGAP is positive in nearly all settings, suggesting that local languages generally provide better access to local cultural knowledge, even when this advantage is hidden by weaker local-language proficiency. Language adaptation reinforces this reading: it moves LOCALGAP and GLOBALGAP consistently, while its effect on KNOWLEDGEGAP is small and inconsistent in sign, indicating that adaptation mainly buys proficiency.

These findings offer useful insight for current localized-knowledge benchmarks and language-adaptation strategies. When a model performs worse in a local language, the cause is often not missing cultural knowledge but weaker proficiency in that language, which masks a knowledge advantage. What is often read as cultural bias or missing cultural knowledge is therefore in large part a multilingual-access problem. Improving proficiency—through language adaptation or stronger multilingual training—is therefore a promising way to unlock this hidden capability, since the knowledge access largely exists and is held back mainly by the proficiency penalty.

Limitations

We note several limitations of this study. First, the size of the culture-specific subsets varies across locales, in part because some languages do not have large-scale cultural benchmark data readily available online.

Second, our evaluation of regionally adapted models is limited in scope. Strong open-weight models that have been explicitly adapted to a non-English language remain relatively scarce, and high-quality, widely-adopted models of this kind are rarer still. So we were able to include only a small number of such models, covering a handful of languages.

Third, all of our results are based on multiple-choice questions scored by log-likelihood. We do not evaluate open-ended generation, where knowledge access may behave differently and where failures can take more diverse forms than a fixed set of options allows.

References

Ishika Agarwal, Nimet Beyza Bozdag, and Dilek Hakkani-Tür. 2025. Language specific knowledge: Do models know better in x than in english? *arXiv preprint arXiv:2505.14990*.

Anthropic. 2026. Claude sonnet 4.6. https://www.anthropic.com/. Large language model developed by Anthropic.

Trevor G. Bond, Zi Yan, and Moritz Heene. 2015. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3 edition. Routledge, New York.

Yu Chen, Telmo Silva Filho, Ricardo B Prudencio, Tom Diethe, and Peter Flach. 2019. β^3 -irt: A new item response model and its applications. In *The 22nd international conference on artificial intelligence and statistics*, pages 1013–1021. PMLR.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier L de Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Google DeepMind. 2025. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Model card PDF, accessed 2026-05-25.

Google DeepMind. 2026. Gemma 4 technical overview and model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-25.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shourya Jain and Paras Chopra. 2026. Language models entangle language and culture. *arXiv preprint arXiv:2601.15337*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. 2025. metabench - a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, volume 2025, pages 31734–31770.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

John P Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL*

2024, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. 2025. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Abdulmohsen Alharthi, Juncai He, Lian Zhang, Haizhou Li, and 1 others. 2024. Alignment at pre-training! towards native alignment for arabic llms. *Advances in Neural Information Processing Systems*, 37:13872–13896.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Fernando Martínez-Plumed, Ricardo B.C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2025. Lost in the mix: Evaluating llm understanding of code-switched text. *arXiv preprint arXiv:2506.14012*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Krzysztof Ociepa, Krzysztof WrĂłbel, Adrian GwoĹşdziej, Remigiusz Kinas, and 1 others. 2024. Bielik 7b v0. 1: A polish language model–development, insights, and evaluation. *arXiv preprint arXiv:2410.18565*.

OpenAI. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, and 1 others. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Lee Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Alfonso Amayuelas, and 1 others. 2025. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pages 83291–83322.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *arXiv preprint arXiv:2401.06466*.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for greek. *arXiv preprint arXiv:2505.13772*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio,

Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2026. Qwen3.5-omni technical report. *Preprint*, arXiv:2604.15804.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek. *arXiv preprint arXiv:2407.20743*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Yang Zhang, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang, and Michalis Vazirgiannis. 2026. Greekmmlu: A native-sourced multitask benchmark for evaluating language models in greek. *arXiv preprint arXiv:2602.05150*.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40:35085–35093.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and 1 others. 2023a. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective.

Yan Zhuang, Qi Liu, Zachary A Pardos, Patrick C Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. 2023b. Position: Ai evaluation should learn from how we test humans. *arXiv preprint arXiv:2306.10512*.

A Model and Dataset Details

We gathered 79 different models with various types, sizes and origins as the population of the benchmarking effort, as further details are shown in Table 6. Both closed-source and open-source models are evaluated, with multiple regionally adapted LLMs considered. In the meantime, both culture-specific (CS) and culture-agnostic (CA) benchmarks are constructed for 14 languages from different language families with spanning richness in resources, as illustrated in Table 7. Multiple languages with extremely low resources are included in particular to study the full spectrum of the knowledge gap of different pretraining scenarios.

B Human evaluation

Annotation Instructions
This project evaluates the cultural ability of LLMs using questions in both the native language and English translation.
You will be given a CSV file containing multiple-choice questions with four options: A, B, C, and D. For each row, please complete two tasks.
Task 1: Translation Quality Score
Evaluate the English translation of the question and all answer options. Enter a score in the SCORE column.
1 – Poor: The translation loses important meaning, mistranslates key or culturally specific terms, or is misleading.
2 – Acceptable: The translation has minor ambiguity or awkwardness, but the overall intent is understandable.
3 – Good: The translation faithfully preserves the original meaning, and the question can be answered correctly from the English version alone.
Task 2: Local Culture Annotation
Indicate whether the question requires local cultural knowledge to answer correctly.
Yes: The question depends on knowledge of the target locale, such as local history, traditions, festivals, geography, laws, institutions, education, literature, religion, food, social norms, or popular culture.
No: The question can be answered using general knowledge and does not rely on local cultural context.
Unsure: The cultural relevance is unclear, weak, or ambiguous.
Please judge each row based on the question and answer options, not on whether you personally know the answer.

Figure 4: The annotation instruction for human native-language-speaking evaluators on translation quality control.

To confirm Assumption 1 for our IRT model, we asked native-speaking translators to verify the quality of the translations for the CS (culture-specific) items. Since every MCQ item originally written in a native language is automatically translated into English, we evaluate each item individually for correctness, rigor, and topic relevance.

For each language that is not the native language among the authors, we hired native speakers to do the annotation, they are paid 15 dollars per hour, and each subset requires around 2 hours of work. Each annotator is given an annotation instruction as described in Figure 4.

Table 8 shows that raw translation quality evaluated by human annotators results in negligible error rates for most local languages (translation scored 1), indicating that translation equivalence is well preserved. Items flagged with a score of 1 or 2 are filtered out, ensuring that the remaining benchmark is fully legitimate.

C Hyperparameters and Setup

The 1PL-IRT model with MLE SGD is trained on an NVIDIA RTX A6000 48GB GPU. The training with learning rate 0.05 spans about 600 to 1000 steps until convergence of the negative log-likelihood loss, taking no more than 1 minute. For ablations, the NUTS fitting process uses 8 sampling chains in parallel, each takes 2000 warmup steps followed by another 2000 sampling steps. For cross validation, all observed responses are partitioned completely randomly into five folds to provide a generalizable held-out cross validation setup.

D Full Specification of the 1PL-IRT Model and Fitting

This appendix gives the complete specification of the facet-conditioned 1PL-IRT model (§3.2) and the fitting procedure (§4.3), including the assumptions, ability parameterization, identifiability argument, and anchoring projection summarized in the main text.

D.1 Model Details

The mathematical validity and interpretability of the facet-conditioned 1PL model rest upon three foundational assumptions concerning the joint distribution of responses and the latent space.

Assumption 4 (Unidimensionality)**.** *The latent trait space is one-dimensional, meaning that a single scalar value θ_{jsk} is strictly sufficient to explain the statistical dependence between item responses for subject j on a subset s with a certain prompt language k.*

Assumption 5 (Local Independence)**.** *Conditional on the latent trait θ_{jsk}, the responses to multi-*

Model Family	Origin	Parameter Sizes	#
CLOSED SOURCE			
GPT-5.4	OpenAI, USA	—	2
Gemini 3 Flash	Google, USA	—	1
Claude Sonnet 4.6	Anthropic, USA	—	1
MULTILINGUAL PRETRAINED			
Llama 3.1	Meta, USA	8B, 70B	4
Llama 3.2	Meta, USA	1B, 3B	4
Gemma 2	Google, USA	9B, 27B	4
Gemma 3	Google, USA	1B, 4B, 12B, 27B	8
Gemma 4	Google, USA	E2B, E4B, 26B-A4B, 31B	8
Qwen 2.5	Alibaba, China	7B, 14B, 32B	6
Qwen 3	Alibaba, China	0.6B, 1.7B, 4B, 14B, 30B-A3B	7
Qwen 3.5	Alibaba, China	0.8B, 2B, 4B, 9B, 27B, 35B-A3B	11
GLM 4 / 4.7	Zhipu AI, China	9B	3
Mistral	Mistral AI, France	119B, 128B	2
Ministral 3	Mistral AI, France	3B, 8B, 14B	6
Jais 2	Inception, UAE	8B, 70B	2
EuroLLM	Europe	9B, 22B	4
Aya Expanse	Cohere, Canada	8B	1
REGIONALLY ADAPTED			
AceGPT-v2	China (Arabic focus)	8B	1
Llama-Krikri	Greece	8B	1
Meltemi	Greece	7B	1
Bielik-Minitron	Poland	7B	1
PersianMind	Iran	—	1
Total			**79**

Table 6: Overview of models evaluated in our benchmark, grouped by family. Both base and instruction-tuned variants are included where available; # denotes the total number of model variants per family.

Language	Code	Family	Resource	#CA	#CS	Source
English	en	Indo-European	5	2,058	6,093	—
Chinese	zh	Sino-Tibetan	4	2,058	1,000	CMMLU
Arabic	ar	Afro-Asiatic	4	2,058	999	ArabicMMLU
French	fr	Indo-European	4	2,058	139	INCLUDE
Japanese	ja	Japonic	4	2,058	73	INCLUDE
Dutch	nl	Indo-European	3	2,058	136	INCLUDE
Hindi	hi	Indo-European	3	2,058	1,000	MILU
Korean	ko	Koreanic	3	2,058	1,000	KMMLU
Persian	fa	Indo-European	3	2,058	185	INCLUDE
Polish	pl	Indo-European	3	2,058	75	INCLUDE
Bengali	bn	Indo-European	2	2,058	342	INCLUDE
Greek	el	Indo-European	2	2,058	785	GreekMMLU
Nepali	ne	Indo-European	1	2,058	230	INCLUDE
Telugu	te	Dravidian	1	2,058	129	INCLUDE
Total	14 languages			**28,812**	**12,186**	

Table 7: Overview of languages and subsets used in our study. The CA (culture-agnostic) subset is sourced from GlobalMMLU. The CS (culture-specific) subset contains locally sourced questions probing culture-specific knowledge; CS items are also available in English translation (CS-en). All subsets are evaluated across 79 LLMs.

Language	Error (%)	Language	Error (%)
Chinese	2.4	Dutch	0.4
Arabic	1.8	Korean	0.9
French	2.0	Greek	0.1
Japanese	4.0	Telugu	0.0
Hindi	1.6	Bengali	2.3
Persian	5.0	Nepali	7.0
Polish	1.7		

Table 8: Error rate (%) in human evaluation of translation quality.

facet problems are mutually independent. For a subset s of size n with prompt language k, the joint probability of a response vector $\boldsymbol{Y}_{jsk} = (y_{1jsk}, y_{2jsk}, \ldots, y_{njsk})$ is the product of the individual marginal probabilities:

$$P(\boldsymbol{Y}_{jsk} = \boldsymbol{y}_{jsk} | \theta_{jsk}, \boldsymbol{b}_s) = \prod_{i=1}^{n} P(Y_{ijsk} = y_{ijsk} | \theta_{jsk}, b_{is}) \quad (6)$$

Assumption 6 (Monotonicity)**.** *The probability of a correct response $P(Y_{ijsk} = 1 | \theta_{jsk}, b_i)$ is a strictly monotonically increasing function of the latent trait θ_{jsk}, meaning that for any two examinees with latent traits $\theta_{1sk}, \theta_{2sk} \in \mathbb{R}$:*

$$\theta_{1sk} < \theta_{2sk} \Leftrightarrow P_i(\theta_{1sk}) < P_i(\theta_{2sk}), \forall i \quad (7)$$

Difficulty-free language contrasts. By treating the language facet as a dimension of the examinee's latent space rather than a shift in item difficulty, the log-odds L_{ijsk} of a correct response for examinee j on item i from subset s in language k is

$$L_{ijsk} = \ln\left(\frac{P(Y_{ijsk} = 1)}{1 - P(Y_{ijsk} = 1)}\right) = \theta_{jsk} - b_i, \quad (8)$$

and likewise $L_{ijsk'} = \theta_{jsk'} - b_i$ for a second language k'. The difficulty term cancels in the within-item difference,

$$L_{ijsk} - L_{ijsk'} = \theta_{jsk} - \theta_{jsk'}, \quad (9)$$

so the language contrast for the same item is exactly the difference in the examinee's language-specific abilities, independent of b_i.

D.2 Fitting Procedure

At our benchmarking scale, classical estimators such as Newton–Raphson or Expectation–Maximization are less practical. We instead frame the facet-conditioned 1PL model as a continuous optimization problem, treating the ability matrix $\boldsymbol{\Theta}$ of all facet-specific examinee abilities and the difficulty vector $\boldsymbol{B}$ of all item difficulties as trainable embeddings and recovering the maximum-likelihood estimates by stochastic gradient descent (SGD).

Let $\boldsymbol{Y} = \{(j, i, s, k, y_{ijsk})\}$ denote the observed benchmarking data, where each entry records examinee j, a multi-facet MCQ item i from subset s prompted in language k, and the dichotomous outcome $y_{ijsk} \in \{0, 1\}$. Writing P_{ijsk} for the logistic probability of Definition 3 and invoking Assumption 5, the joint likelihood is the product of the individual Bernoulli probabilities:

$$L(\boldsymbol{\Theta}, \boldsymbol{B} \mid \boldsymbol{Y}) = \prod_{(j,i,s,k,y_{ijsk}) \in \boldsymbol{Y}} P_{ijsk}^{y_{ijsk}} (1 - P_{ijsk})^{1 - y_{ijsk}}. \quad (10)$$

Taking the natural logarithm yields the log-likelihood ℓ, and the empirical loss $\mathcal{L}$ is the negative log-likelihood (NLL), which coincides with the binary cross-entropy (BCE) loss of standard machine-learning toolkits:

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{B}) = -\ell(\boldsymbol{\Theta}, \boldsymbol{B}) = -\log L(\boldsymbol{\Theta}, \boldsymbol{B} \mid \boldsymbol{Y}). \quad (11)$$

We minimize $\mathcal{L}$ over $\boldsymbol{\Theta}$ and $\boldsymbol{B}$ jointly using the Adam optimizer (Kingma and Ba, 2017) with mini-batch SGD.

Identifiability and anchoring. The 1PL model is non-identifiable: because the logistic function depends only on the difference $\theta_{jsk} - b_i$, adding any constant c to all abilities and all difficulties leaves every probability unchanged,

$$\sigma\big((\theta_{jsk} + c) - (b_i + c)\big) = \sigma(\theta_{jsk} - b_i). \quad (12)$$

To guarantee a unique MLE, we anchor the latent scale by constraining the mean item difficulty to zero. After each SGD update we apply the projection, with $\bar{b} = \frac{1}{|I|} \sum_{i \in I} b_i$,

$$\begin{aligned} b_i &\leftarrow b_i - \bar{b}, \quad \forall i \in I, \\ \theta_{jsk} &\leftarrow \theta_{jsk} - \bar{b}, \quad \forall j, s, k. \end{aligned} \quad (13)$$

This re-centering fixes the origin of the latent scale and gives a unique, interpretable solution.

D.3 IRT Variant Details

Item-response representations. The **dichotomous** (binary) model scores each response as correct or incorrect ($y \in \{0, 1\}$) and models $P_i(\theta_j) =$

$\sigma(\theta_j - b_i)$ as in the main text (§6.4). The **Continuous Response Model (CRM)** instead treats the response as graded: rather than thresholding, it uses the probability mass the model places on the correct option after normalizing over the answer-option tokens, giving a continuous score in $[0, 1]$. This retains finer information about how confident the model was, at the cost of inheriting any miscalibration in the raw token probabilities.

Inference algorithms. **MLE** treats θ and b as free parameters with no prior and minimizes the negative log-likelihood by gradient descent (Adam), returning a single point estimate $(\hat{\theta}, \hat{b})$. **NUTS** is an adaptive variant of Hamiltonian Monte Carlo that draws samples from the full Bayesian posterior over (θ, b); it is a standard choice for fitting IRT models and yields credible intervals directly, but its cost grows with the dimensionality of the parameter space and it can suffer from divergences and inefficient exploration in very high-dimensional settings.

E Full results

We present results for all models tested in Figure 5.

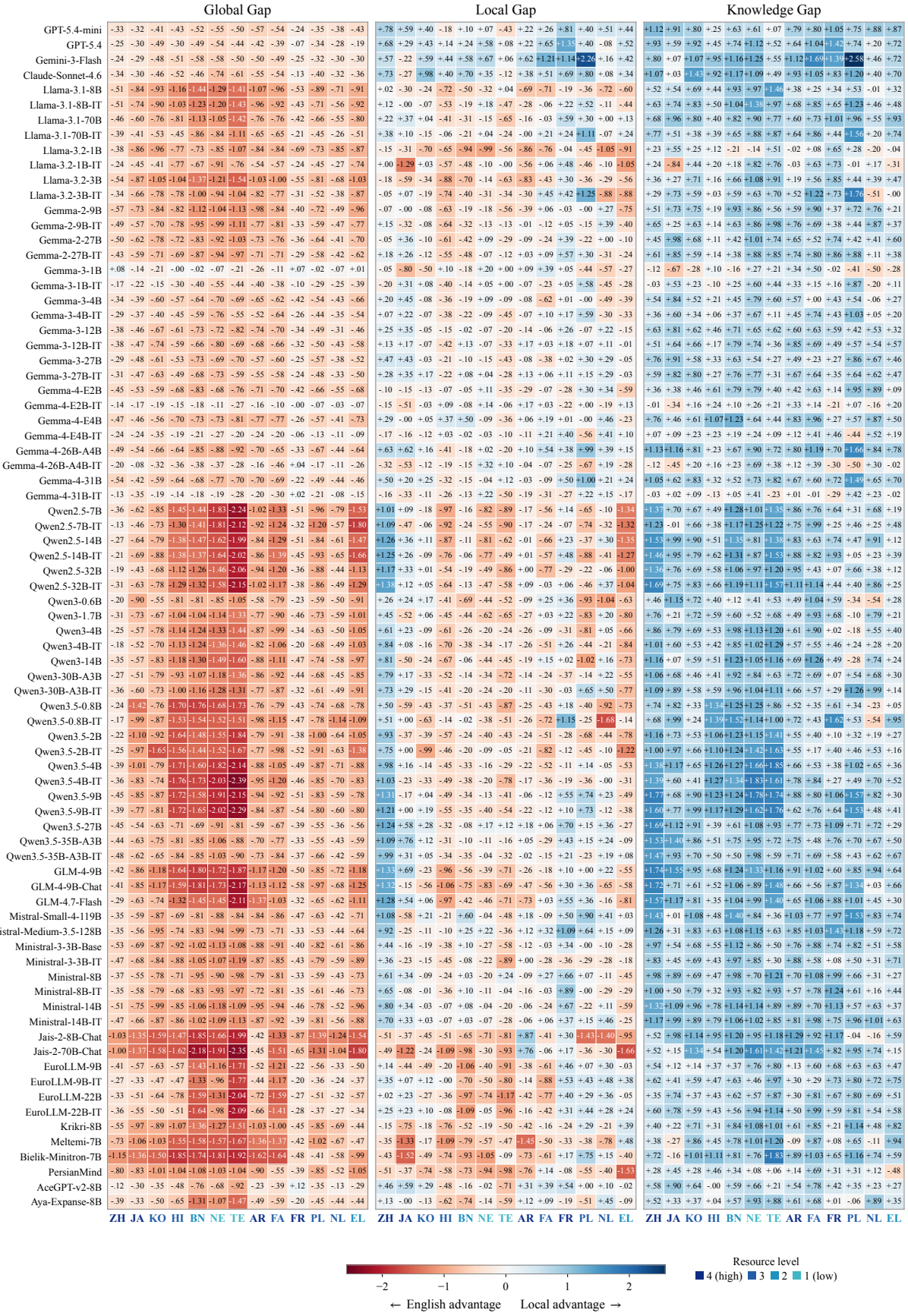

Figure 5: GlobalGap, LocalGap, and KnowledgeGap across all models and locales.